POWER ESTIMATES IN RANDOMIZED TRIALS FOR MISTIE II/III BASED ON SIMULATIONS ASSUMING A DICHOTOMOUS/ORDINAL OUTCOME

by
Yuying Yan

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
May, 2021

Abstract

Intracerebral hemorrhage (ICH) accounts for 2 million (10-15%) of all 15 million strokes worldwide each year. Over the last decades, many research trials have sought to understand ICH, improve patients' functional outcomes, and lower mortality. This thesis presents power analyses based on simulating data from the Minimally Invasive Surgery (MIS) with thrombolysis in a Phase 2 (MISTIE II) and a Phase 3 (MISTIE III) trial. We introduce different scenarios in these simulations. The trial designs consist of four main parts: (1) binary endpoint versus ordinal endpoint, (2) futility stopping rules, (3) MIS evacuation timing, and (4) patients' clot size at the end of surgery. This work aims to provide the appropriate rationale for MIS training in future studies. From the results, we suggest future clinical studies should have as many patients as possible achieving 15cc or less clot size at the end of MIS. It is reasonable if the sample size of each arm $\geq$ 800 with power = 80%. If clinicians prefer a smaller sample size, then analysis of treatment effects with covariate adjustment and ordinal endpoints can be considered.

Primary Reader and Advisor: Dr. Richard E. Thompson

Secondary Reader: Dr. Michael Rosenblum

Acknowledgments

I am grateful to my advisor, Dr. Richard E. Thompson, for introducing me to the fascinating research topic and guiding me through my ScM program. I appreciate Dr. Michael Rosenblum's help and suggestions for this research topic. It is my great honor to acknowledge Dr. Daniel Hanley and the Brain Injury Outcomes for allowing me to use the MISTIE data. I am also grateful to Dr. Elizabeth Colantuoni for always kindly supporting our cohort and selflessly providing help. I would like to express my gratitude to the Biostatistics department for the guidance along the road of my life in graduate school. I couldn't have such fantastic graduate life without the department's extraordinary environment.

Moreover, I am extremely thankful to my parents for their trust and support. Their love embraces me to be a better person. Thank you for all my friends supporting me when I was depressed and hesitated for my future.

I hope everyone can have a bright future.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Intracerebral hemorrhage (ICH) is a disastrous event caused by ruptured vessels affected by hypertension-related degenerative changes or cerebral amyloid angiopathy.[1] ICH accounts for 2 million (10-15%) of all 15 million strokes worldwide each year.[2] It has a very high morbidity and mortality rate that has not changed over the last 30 years. [3] Unfortunately, there is still no evidence-based primary treatment for this condition.[4][5]

Over the last decades, many research trials have sought to understand ICH, improve functional outcomes, and lower mortality. Since the modified Rankin Scale (mRS), a 7-level ordered categorical score ranging from 0 (no symptoms at all) to 6 (dead), has been regarded as a valid and reliable endpoint in randomized clinical trials, it is recommended to represent outcomes for stroke studies in estimating the degree of disability for patients who have had a stroke. [6][7][8] The Minimally Invasive Surgery (MIS) with thrombolysis in a Phase 2 trial (MISTIE II) and Phase 3 (MISTIE III) trial both used the mRS to measure patients' functional outcome. MISTIE II shows the advantage of patients' functional outcomes at interim analysis 180 days.[9] Exploratory analysis in MISTIE III mentions an association between the extent of clot removal and lower mRS scores. It points out the need for an additional clinical trial in estimating the

effect of this therapy. [10][11] In particular, more investigation needs to be performed on the timing and optimal performance of MIS. Therefore, a power analysis based on MISTIE II/III and more rigorous MIS performance criteria are important in providing an appropriate rationale for training and clinical experience in the future.

1.2 Statistical background

When researchers plan a study, the required sample size is always a big concern. To provide an appropriate sample size, evaluating the statistical power based on simulation is one strategy. In this simulated study, power is the probability of finding a statistical difference between MISTIE and standard medical treatment performance in patients' functional outcomes when the average treatment effect is the minimum, clinically meaningful treatment effect.

We also consider the impact of a futility analysis. Early futility stops can minimize patients' exposure to ineffective treatments and save the use of resources. Conditional power is a popular strategy to examine the futility of a trial. In addition, there exist meta-analysis studies suggesting that beneficial effects remain true when analyzing evacuation timing subgroups. [12] Therefore, in this simulated study, we would like to examine the statistical difference in treatment effect considering the futility.

Our main contribution is to simulate a randomized clinical ICH trial using data derived from MISTIE II/III. We estimate statistical power under the case of a binary endpoint versus ordinal endpoint for various sample sizes, consider the probability of removing one or both of two different surgery intervention arms due to futility, and evaluate and quantify various trial results at the completion of simulated patient recruitment.

1.3 Organization of This Article

Chapter 2 introduces the demographics and clinical characteristics with results in binary endpoint versus ordinal endpoint based on MISTIE II/III patients. Chapter 3 compares the statistical power under two endpoint scenarios (binary and ordinal outcomes). Chapter 4 further evaluates the statistical power using simulated data while considering futility and multiple comparisons. Chapter 5 is devoted to a discussion of the results.

Chapter 2

Basic information

The data are derived from two sources: MISTIE II (clinicaltrials.gov NCT00224770), a multicenter, randomized, and open-label phase 2 trial; MISTIE III (clinicaltrials.gov NCT01827046), a multicenter randomized, controlled, open-label, blinded endpoint phase 3 trial. All patients are over the age of 18 and have a Historical Rankin score of 0 or 1. Participants have a Glasgow Coma Scale (GCS) 14 or less or a National Institute of Health Stroke Scale 6 or more at the time of presentation.[9][10] Patient functional outcome is estimated by modified Rankin Scale(mRS) score.

2.1 Basic Demographic and Clinical Characteristics

Demographic and baseline characteristics for the whole dataset can be found in table 2.1. “Stability CT ICH” denotes Stability CT Intracerebral hemorrhage(ICH) volume, “Stability CT IVH” denotes Stability CT Intraventricular hemorrhage (IVH) volume, Glasgow Coma Scale (GCS) is a scoring system used to describe patients’ degree of consciousness following a brain injury. [14] The clot location is either lobar or deep (putamen or thalamus).

	Medical (N=293)	Surgical (N=347)	Overall (N=640)
age			
Mean (SD)	61.3 (12.8)	61.0 (11.6)	61.2 (12.2)
Median [Min, Max]	62.0 [28.0, 90.0]	62.0 [29.0, 84.0]	62.0 [28.0, 90.0]
Missing	0 (0%)	39 (11.2%)	39 (6.1%)
Glasgow Coma Scale			
3-8	76 (25.9%)	98 (28.2%)	174 (27.2%)
9-12	121 (41.3%)	145 (41.8%)	266 (41.6%)
13-15	96 (32.8%)	101 (29.1%)	197 (30.8%)
Missing	0 (0%)	3 (0.9%)	3 (0.5%)
clot location			
Deep(putamen or thalamus)	172 (58.7%)	230 (66.3%)	402 (62.8%)
Lobar	121 (41.3%)	117 (33.7%)	238 (37.2%)
stability CT ICH			
Mean (SD)	47.6 (17.5)	48.2 (18.7)	47.9 (18.1)
Median [Min, Max]	44.5 [16.8, 119]	44.8 [15.0, 127]	44.6 [15.0, 127]
stability CT IVH			
Mean (SD)	2.62 (4.70)	3.05 (6.45)	2.85 (5.71)
Median [Min, Max]	0.500 [0, 29.5]	0.330 [0, 61.8]	0.400 [0, 61.8]
mRS at 180 days(ordinal)			
0	4 (1.4%)	2 (0.6%)	6 (0.9%)
1	5 (1.7%)	8 (2.3%)	13 (2.0%)
2	25 (8.5%)	36 (10.4%)	61 (9.5%)
3	69 (23.5%)	77 (22.2%)	146 (22.8%)
4	67 (22.9%)	79 (22.8%)	146 (22.8%)
5	43 (14.7%)	60 (17.3%)	103 (16.1%)
6	69 (23.5%)	55 (15.9%)	124 (19.4%)
Missing	11 (3.8%)	30 (8.6%)	41 (6.4%)
mRS at 180 days(binary)			
0	179 (61.1%)	194 (55.9%)	373 (58.3%)
1	103 (35.2%)	123 (35.4%)	226 (35.3%)
Missing	11 (3.8%)	30 (8.6%)	41 (6.4%)

Table2.1 Demographic and baseline characteristics for MISTIE II/III, “mRS”: Patients Modified Rankin Scale(mRS) - ordinal at day 180, range from 0 to 6 where 0 is no symptom, 1 is no significant disability and able to carry out all pre-stroke activities, 2 is slight disability and unable to carry out all pre-stroke activities, 3 is moderate disability with requiring some external help, 4 is moderately severe disability and unable to walk or attend to bodily functions without others’ assistance, 5 is severe disability with requiring continuous care,6 is death; binary at day 180, 1 = good mRS (0-3), 0 = bad mRS (4-6)

2.2 Model Results

2.2.1. MISTIE vs. Medical

We perform statistical analysis with the completed data (581 patients) using the generalized logistic regression model and generalized ordinal logistic regression model. For the generalized logistic regression model, the outcome is defined as the proportion of patients who achieved an mRS score of 0-3 at 180 days; for the generalized ordinal logistic regression model, the ordinal outcome is defined as the proportion of patients who achieved an mRS score of 0-2, 3, 4, 5, 6 at 180 days. Differences between MIS and medical group with adjustment in baseline covariates (stability intracerebral hemorrhage size, stability intraventricular hemorrhage size, age, GCS, clot location) are examined under these two scenarios. The result is shown in Table 2.2.1.1.

	MIS vs. Medical (Binary Endpoint)		
Predictors	*Odds Ratios*	*Statistic*	*p*
MIS	1.42	1.66	0.098
age	0.91	-8.30	**<0.001**
GCS 3-8	0.23	-4.87	**<0.001**
GCS 9-12	0.42	-3.63	**<0.001**
Location-Globus Lobar	10.97	8.57	**<0.001**
stability CT ICH	0.96	-5.12	**<0.001**
stability CT IVH	0.96	-1.62	0.104

	MIS vs. Medical (Ordinal Endpoint)	
Predictors	*Statistic*	*p*
MIS	-2.45	**0.014**
GCS 3-8	6.28	**<0.001**
GCS 9-12	2.78	**0.005**
stability CT ICH	7.10	**<0.001**
stability CT IVH	4.43	**<0.001**

	MIS vs. Medical (Ordinal Endpoint) nominal variables	
Predictors	*Statistic*	*p*
0-2|3 age	-4.76	**<0.001**
3|4 age	-9.05	**<0.001**
4|5 age	-7.55	**<0.001**
5|6 age	-4.11	**<0.001**
0-2|3 Lobar	6.59	**<0.001**
3|4 Lobar	9.07	**<0.001**
4|5 Lobar	5.06	**<0.001**
5|6 Lobar	-0.04	0.969

Table 2.2.1.1 Comparsion of the binary endpoint versus ordinary endpoint model results with covariate adjustment. In ordinary endpoint, nominal specifies variables which are not proportional to the outcome

2.2.2 MIS vs. Medical regarding patients symptom onset to surgery hours

Regarding the time duration from patients' symptoms onset to surgery time, most patients received surgery between 40 to 75 hours after symptom onset (Figure 2.2.2.1). Patients were separated into two groups, early surgery and late surgery groups, by median (55 hours) of symptoms onset to surgery. Demographic and baseline characteristics are shown in Appendix 1.

Statistical analysis is performed by using the generalized logistic regression model and generalized ordinal logistic regression model. The outcome is defined as the same in 2.2a for both models. Treatment differences between each of the three groups (late, early, medical) with adjustment for baseline covariates (stability intracerebral hemorrhage size, stability intraventricular hemorrhage size, age, GCS, clot location) are examined. The result is shown in table 2.2.2.1 and table 2.2.2.2.

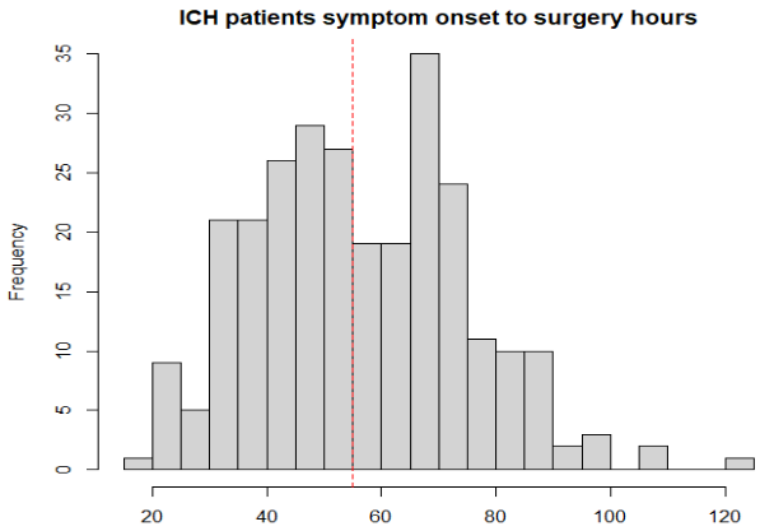


Figure 2.2.2.1 Distribution of ICH patients' symptom onset to MIS surgery time. The red dashed line denotes median hours across patients in the surgery group

	Early (n = 136)			Late (n = 139)		
Predictors	*Odds Ratios*	*Statistic*	*p*	*Odds Ratios*	*Statistic*	*p*
MIS	1.36	1.16	0.245	1.59	1.70	0.089
age	0.91	-6.85	**<0.001**	0.91	-7.31	**<0.001**
GCS 9-12	2.04	2.12	**0.034**	1.57	1.34	0.182
GCS 13-15	3.94	3.83	**<0.001**	4.18	3.93	**<0.001**
Location-Lobar	8.67	6.67	**<0.001**	13.03	7.52	**<0.001**
stability CT ICH	0.97	-3.95	**<0.001**	0.95	-5.27	**<0.001**
stability CT IVH	0.98	-0.74	0.457	0.99	-0.24	0.811

Table 2.2.2.1 Binary endpoint model results. Left: Early Surgery patients (n=136) versus standard medical patients; Right: Late Surgery patients (n=139) versus Standard medical patients

	Early (n = 136)			Late (n = 139)		
Predictors	*Odds Ratios*	*Statistic*	*p*	*Odds Ratios*	*Statistic*	*p*
MIS	0.74	-1.48	0.138	0.61	-2.56	**0.011**
GCS 9-12	1.03	4.55	**<0.001**	1.04	6.06	**<0.001**
GCS 13-15	1.04	2.32	**0.020**	1.02	0.81	0.417
stability CT ICH	0.44	-3.50	**<0.001**	0.50	-3.01	**0.003**
stability CT IVH	0.24	-5.26	**<0.001**	0.20	-6.04	**<0.001**
nominal variables						
Predictors	*Odds Ratios*	*Statistic*	*p*	*Odds Ratios*	*Statistic*	*p*
0-2|3 age	0.93	-4.50	**<0.001**	0.95	-3.65	**<0.001**
3|4 age	0.92	-7.39	**<0.001**	0.91	-8.17	**<0.001**
4|5 age	0.93	-6.56	**<0.001**	0.93	-6.35	**<0.001**
5|6 age	0.96	-3.57	**<0.001**	0.97	-2.85	**0.004**
0-2|3 Lobar	10.16	5.83	**<0.001**	7.45	5.43	**<0.001**
3|4 Lobar	7.42	7.04	**<0.001**	10.16	8.11	**<0.001**
4|5 Lobar	2.86	3.90	**<0.001**	3.24	4.46	**<0.001**
5|6 Lobar	0.99	-0.03	0.978	1.02	0.05	0.958

Table 2.2.2.2 Ordinal endpoint model results. Left: Early Surgery patients (n=136) versus Standard medical patients; Right: Late Surgery patients(n=139) versus standard medical patients; nominal specifies variables which are not proportional to the outcome

Chapter 3

MIS and standard medical binary versus the ordinal outcome

3.1 Method

This chapter aims to estimate and compare statistical power under two scenarios: (1) Binary endpoint via generalized logistic regression model; and (2) Ordinal endpoint via generalized ordinal logistic regression model. We also explore the impact of higher percentage participants achieving clot size 15cc or less at the end of treatment in computing statistical power. Both scenarios are performed without covariate adjustment. In each scenario, we simulate MIS and medical groups with 1:1 randomization to the two arms with total enrollments of n = 350, 500, 600, 800, and 1000. We generate data for standard medical-arm participants by resampling with replacement from the original control group of MISTIE data. The "no enriched" MIS arm participants are generated by resampling with replacement from the initial surgery group of MISTIE data. Data for "enriched" MIS-arm participants are simulated in two steps: (1) Dividing original MISTIE data by patients clot size ≤ 15 and > 15 cc at the end of treatment. (2) Participants who reach clot sizes less than 15 and greater than 15 cc are resampled with replacement from each stratum to reach prespecified ratios = 9:1, 8.5:1.5, 8:2, 7:3. We compare the power of binary endpoint to the ordinal endpoint with different percentages of patients

having 15cc or less clot at the end of treatment in each simulated MISTIE arm. These endpoints are defined as below and implemented in the R studio.

3.2 Binary and Ordinal Endpoint

In binary endpoint, Odds Ratio is the ratio of odds of good functional outcome (mRS 0-3) comparing the MIS surgery group to the standard medical group.

Odds(Y = 1| A = 1)/Odds (Y=1|A = 0) Y: functional outcome, A: group{1: surgery, 0: medical}

For ordinal endpoint, its outcome has 5 levels: mRS0-2(level1), mRS 3(level 2), mRS 4(level 3), mRS 5(level 4), mRS 6(level 5).

Cumulative log Odds Ratio: cumulative log odds ratio of functional outcome at level 1 to level k comparing the MIS surgery group to the standard medical group.

3.3 Performance criteria

We compare the type I error with significance level 0.05 and power of test regarding the null hypothesis H_0: no effect of MIS treatment compared to standard of care (medical treatment)

3.4 Simulation Results

N	Outcome Type	No Enhancement P(reject $\boldsymbol{H_0}$)	70% Enhancement P(reject $\boldsymbol{H_0}$)	80% Enhancement P(reject $\boldsymbol{H_0}$)	85% Enhancement P(reject $\boldsymbol{H_0}$)	90% Enhancement P(reject $\boldsymbol{H_0}$)
350	Binary	0.097	0.345	0.543	0.663	0.768
500	Binary	0.132	0.439	0.722	0.815	0.902
600	Binary	0.155	0.512	0.787	0.872	0.939
800	Binary	0.210	0.643	0.888	0.952	0.971
1000	Binary	0.250	0.765	0.955	0.979	0.995

350	Ordinal	0.371	0.732	0.884	0.946	0.961
500	Ordinal	0.497	0.859	0.972	0.990	0.997
600	Ordinal	0.581	0.916	0.987	0.997	1
800	Ordinal	0.701	0.966	1	1	1
1000	Ordinal	0.796	0.994	0.999	1	1

Table3.4.1 Simulated power results for the binary outcome and ordinal outcome under the different scenarios. N denotes simulated arm size for the MIS and the standard medical groups; "Enhancement" denotes the proportion of patients having 15cc or less clot at the end of treatment in simulated MIS arm.

3.5 Discussion

In Table 3.4.1, the first five rows compare the performance of the binary outcomes in different treatment arm sizes where the good functional outcome is defined as patients achieve mRS 0-3 at 180 days. After considering different enhancement proportions, absolute gains in power vary from 15% to 42% when arm size increases from 350 to 1000. As seen in Table 3.4.1, the last five rows compare the ordinal outcomes in the different arms. The covariates-adjusted method achieves higher power when arm size becomes larger. After considering different enrichment proportions, absolute gains in power vary from 4% to 42% when arm size increases from 350 to 1000. Overall, Table 3.4.1 shows that performing ordinal endpoint and higher percentage enhancement in patients with 15cc or less clot size at the end of treatment can achieve higher power than binary endpoint across all settings. Absolute gains in power vary from 4% to 42%. Under some scenarios, power reaches 100%, and we conclude it is a possible overpower issue and should consider whether we waste clinical resources in enrolling unnecessary patients. Chapter 4 will discuss those issues in more detail.

Chapter 4

Interim analysis in MIS Arm versus standard medical arm with the binary outcome

4.1 Introduction

This chapter aims to compare various trial results when considering different percentage participants achieving clot size 15cc or less, evacuation timing, and the probability of dropping MIS intervention arms due to futility. The resulting multiple comparison problem is also taken into account. Late phase trials often recruit hundreds or thousands of participants. It is reasonable to consider whether the trial should be stopped earlier for futility when examining interim data, avoiding wasting clinical resources, and avoiding giving patients ineffective treatments.[14] One of the main methods for futility analysis is based on conditional power. It helps calculate the probability of obtaining a final significant result conditional on the data obtained at the interim analysis. If the probability is below a predefined threshold, the trial is terminated early. The planned binary outcome is defined as a good functional outcome (mRS 0-3 at 180 days) comparing MIS surgery groups of early (evacuation timing < 55 hrs) and late (evacuation timing ≥ 55hrs) to the standard medical group.

4.2 Statistical Framework

4.2.1 Conditional Power

Jennison and Turnbull (2000) illustrate the general idea in calculating conditional power [15]:

The general upper one-sided conditional power at stage k for rejecting a null hypothesis about a parameter θ at the end of the study, given the observed test statistics, Z_k, is computed as

$$P_{uk}(\theta) = \Phi\left(\frac{Z_k\sqrt{I_k} - z_{1-\alpha}\sqrt{I_K} + \theta(I_K - I_k)}{\sqrt{I_K - I_k}}\right)$$

where

θ = the parameter tested by the hypothesis at the end of the study
k = an interim stage at which the conditional power is computed
K = the stage that final test is computed
Z_k = the test statistic calculated from data obtained at an interim stage so far
I_k = the information level at stage k
I_K = the information level at the end of the study
$Z_{1-\alpha}$ = z-statistic for the test with a type I error rate of α

Consider a test of proportions of patients having mRS 0-3 at 180 days, with the following hypothesis, where P_{Med} and P_{Surg} are the population proportions in the MIS group and the standard medical group.

$$H_0: P_{Med} = P_{Surg}$$

$$H_a: P_{Med} < P_{Surg}$$

Components in computing conditional power for the proportion test are in Chang(2008)[16]

$\theta = P_{Surg} - P_{Med}$ the expected difference under the alternative hypothesis

$Z_k = \left(p_{Surg} - p_{Med}\right)\sqrt{\widehat{I_k}}$

(the z-statistic from the data obtained at the interim stage so far)

$I_k = \frac{1}{\sigma^2}\left(\frac{1}{n_{Surg}} + \frac{1}{n_{Med}}\right)^{-1}$ the interim information level

$I_K = \frac{1}{\sigma^2}\left(\frac{1}{N_{Surg}} + \frac{1}{N_{Med}}\right)^{-1}$ the final information level

where

$\mathrm{p_{Surg}} =$ sample proportion for the MISTIE group at an interim stage

$\mathrm{p_{Med}} =$ sample proportion for the standard medical group at an interim stage

$\widehat{I_k} =$ estimated information from the sample at an interim stage

$n_{\mathrm{Surg}} =$ MISTIE group sample size at an interim stage

$n_{\mathrm{Med}} =$ Standard medical group sample size at an interim stage

$N_{\mathrm{Surg}} =$ MISTIE group final sample size

$N_{\mathrm{Med}} =$ Standard medical group final sample size

$\sigma^2 = \overline{\mathrm{p}}(1 - \overline{\mathrm{p}})$ with $\overline{p} = \frac{\left(P_{Surg} + P_{Med}\right)}{2}$

Note that $\mathrm{P_{Med}}$ and $\mathrm{P_{Surg}}$ should be prespecified

Generally, if $P_{uk}(\theta)$ falls below 0.2 or 0.1, the study may be stopped.

4.2.2 Bonferroni Correction

Goeman and Solari (2014) review the general idea of Bonferroni correction for familywise error control on multiple testing [17]:

Let $\mathrm{H_1}, \ldots \mathrm{H_m}$ be a family of hypotheses and $p_1, \ldots p_m$ be their corresponding p-values. Let m be the total number of null hypotheses and m_0 be the number of the true null hypothesis. The familywise error rate (FWER) is the probability of rejecting at least one true H_i. The Bonferroni correction rejects the null hypothesis for each $\mathrm{p_i} \leq \frac{\alpha}{\mathrm{m}}$, therefore controlling the FWER at $\leq \alpha$.

$$\text{FWER} = \text{P} \cup_{\text{i}=1}^{\text{m}_0} \left(\text{p}_\text{i} \leq \frac{\alpha}{\text{m}}\right) \leq \sum_{i=1}^{m_0} P\left(p_i \leq \frac{\alpha}{m}\right) = \text{m}_0 \frac{\alpha}{m} \leq m \frac{\alpha}{m} = \alpha$$

4.3 Method

We simulate the MIS early surgery group (evacuation timing < 55 hrs), MIS late surgery group (evacuation timing ≥ 55 hrs), and medical group with 1:1:1 randomization to the three arms with a total enrollment of n = 350, 500, 600, 800 and 1000. We generate data for standard medical-arm participants by resampling with replacement from the original control group of MISTIE data. The "no enriched" MIS arm participants are generated by resampling with replacement from the original surgery group of MISTIE data. Data for "enriched" MIS-arm participants are simulated in three steps: (1) Dividing MISTIE data by patients' median evacuation timing into two stratums, (2) In each stratum, dividing data by patients clot size less than 15 and greater than 15 cc at the end of treatment. (3) Participants with clot size less than 15 and greater than 15 cc are resampled with replacement from each group in the prespecified ratios = 9:1, 8.5:1.5, 8:2, 7:3. In each case, 1000 trials are simulated. We compute conditional power to decide if any arm should be dropped at the interim analysis (i.e., simulating 1/2 of the total enrollment of n patients data) based on the expected difference under the alternative hypothesis at the end of the study. The expected difference is determined by the original MIS date and power we expect to see at the endpoint. We assume true treatment effects are the same in MIS early vs. Medical, MIS late vs. Medical. When both surgery arms are dropped, the entire study will halt, and we don't proceed to power calculation at the end of the study. We tabulate results for the following scenarios: (1) testing the endpoints and calculating power using binary outcome with one-sided level and Bonferroni Correction since multiple endpoints, (2) different scale in expected difference for the percentage of patients having mRS 0-3 at 180 days between

MISTIE(late/early) arm and standard medical arm, (3) conditional power with 0.1 and 0.2 cutoffs. All simulations are implemented in R studio.

4.4 Results

4.4.1 Baseline

Table 4.4.1.1 shows the percentages of patients having mRS 0-3 at 180 days from the original patient data in MISTIE II/III.

$P_{\mathbf{Med}}$	P_{Surg_Early}	P_{Surg_Late}
36.5%	35.5%	43.2%

Table 4.4.1.1 Proportion of patients having mRS 0-3 at 180 days from the original MISTIE II/III patient data. "Med" denotes patients who received standard medical care; "Surg_Early" denotes patients who received MIS with evacuation timing less than 55 hours; "Surg_Late" denotes patients who received MIS with evacuation timing larger than 55 hours

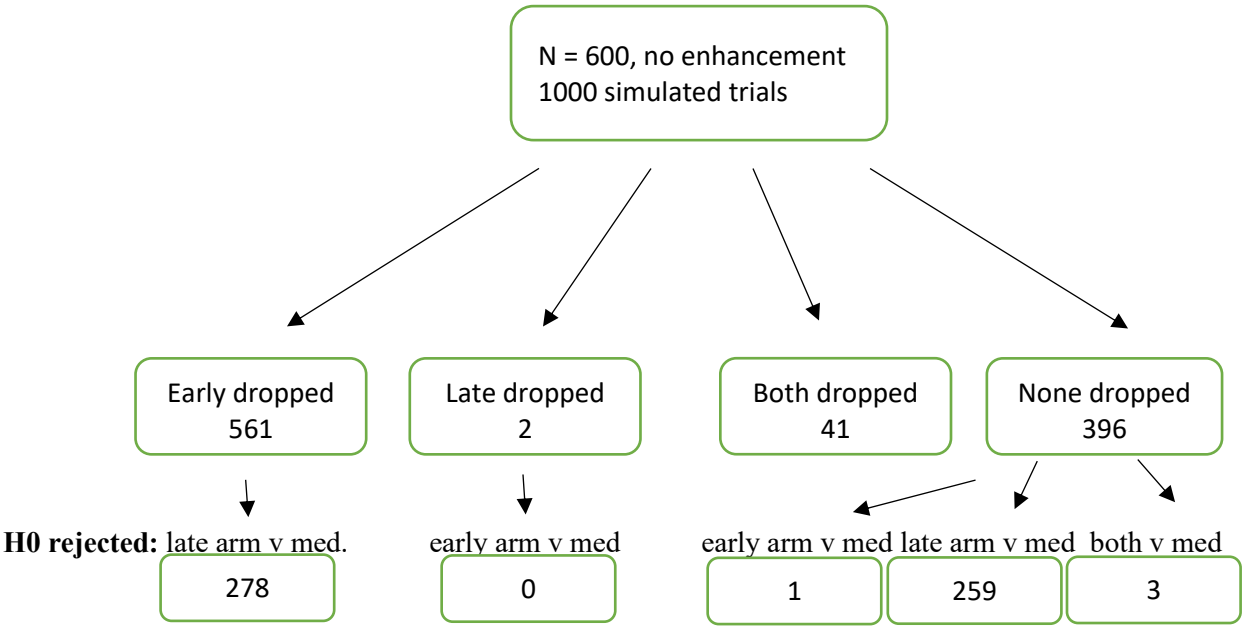
4.4.2 Expected Difference, Power and Futility Cutoff

N	$\boldsymbol{\alpha}$ Type I error rate	Expected Power	$P_{\mathbf{Med}}$	$\boldsymbol{\theta}$	Stopping Cutoff
350	0.025	80%	0.365	0.1045	0.1/0.2
500	0.025	80%	0.365	0.087	0.1/0.2
600	0.025	80%	0.365	0.0795	0.1/0.2
800	0.025	80%	0.365	0.069	0.1/0.2
1000	0.025	80%	0.365	0.062	0.1/0.2
350	0.025	90%	0.365	0.121	0.1/0.2
500	0.025	90%	0.365	0.101	0.1/0.2
600	0.025	90%	0.365	0.092	0.1/0.2
800	0.025	90%	0.365	0.0795	0.1/0.2
1000	0.025	90%	0.365	0.071	0.1/0.2

Table 4.4.2.1 Proposed different scenarios that are estimated at the interim analysis. P_{Med} is the prespecified value for the population proportion of patient achieving mRS 0-3 at 180 days for standard medical arm based on the trial data; θ denotes the expected minimally detectable difference of population proportion between the surgery arm (P_{Surg}) and the standard medical arm (P_{Med}) under different desired power; and "Stopping cutoff" denotes thereshold for dropping arm at the interim analysis.

Table 4.4.2.1 shows the expected difference (θ) of population proportion of patients having mRS 0-3 at 180 days between the MIS (late/early) arm and the standard medical arm under different cases of desired power, the sample size of each arm, and type I error rate. We use θ from the table to calculate the conditional power in the simulated studies and determine whether the MIS arm(s) should be dropped or not given prespecified futility criteria.

4.4.3 Results



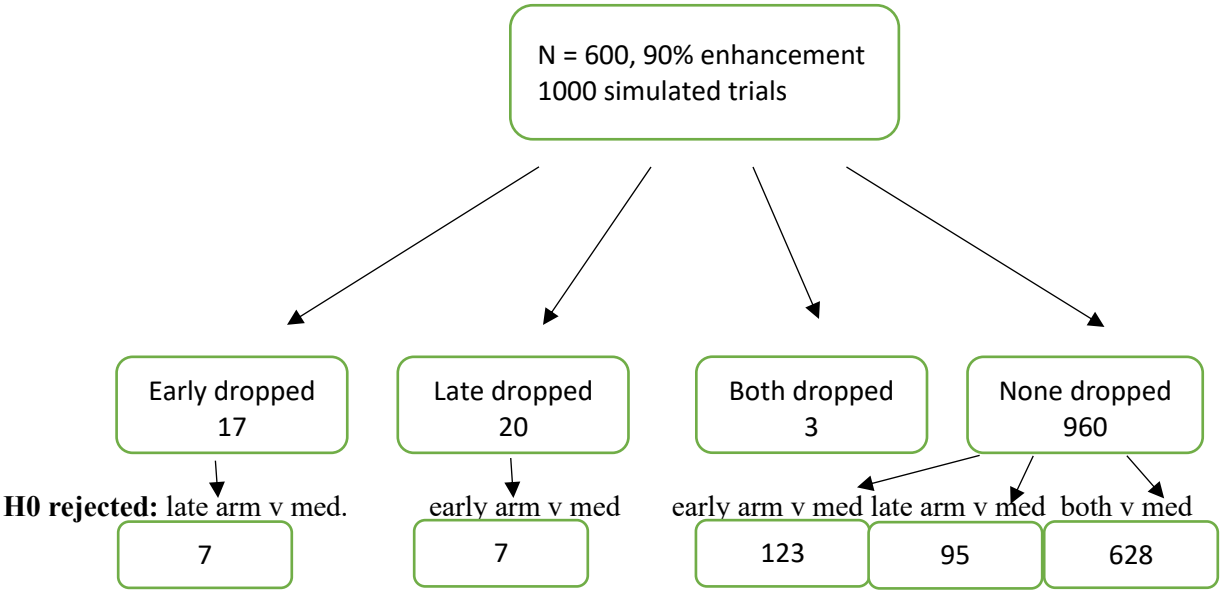


Figure 4.4.3.1 Futility drop results and the number of simulated trials detecting treatment effects. The sample size of each arm is n=600, interim stopping criteria is 0.2 and the enhancement is prespecified

Figure 4.4.3.1 shows the results when the sample size of each arm is 600. The second row of each flow chart shows the futility drop results between simulated MIS arm having no enhancement and arm having 90% patients down to 15cc or less clot at the end of treatment. The stopping cutoff at interim analysis is specified to 0.2, and the desired level of true treatment effect is 0.0795 comparing MIS (early/late) arm vs. Medical arm. The third row of each flow chart represents the number of trials detecting the difference between the surgery arm and standard medical arm under different futility scenarios (early/late/both surgery arms dropped at the interim analysis). Detailed futility results with different cases of power, enhancement fraction, stopping cutoff are displayed in Appendix 2.

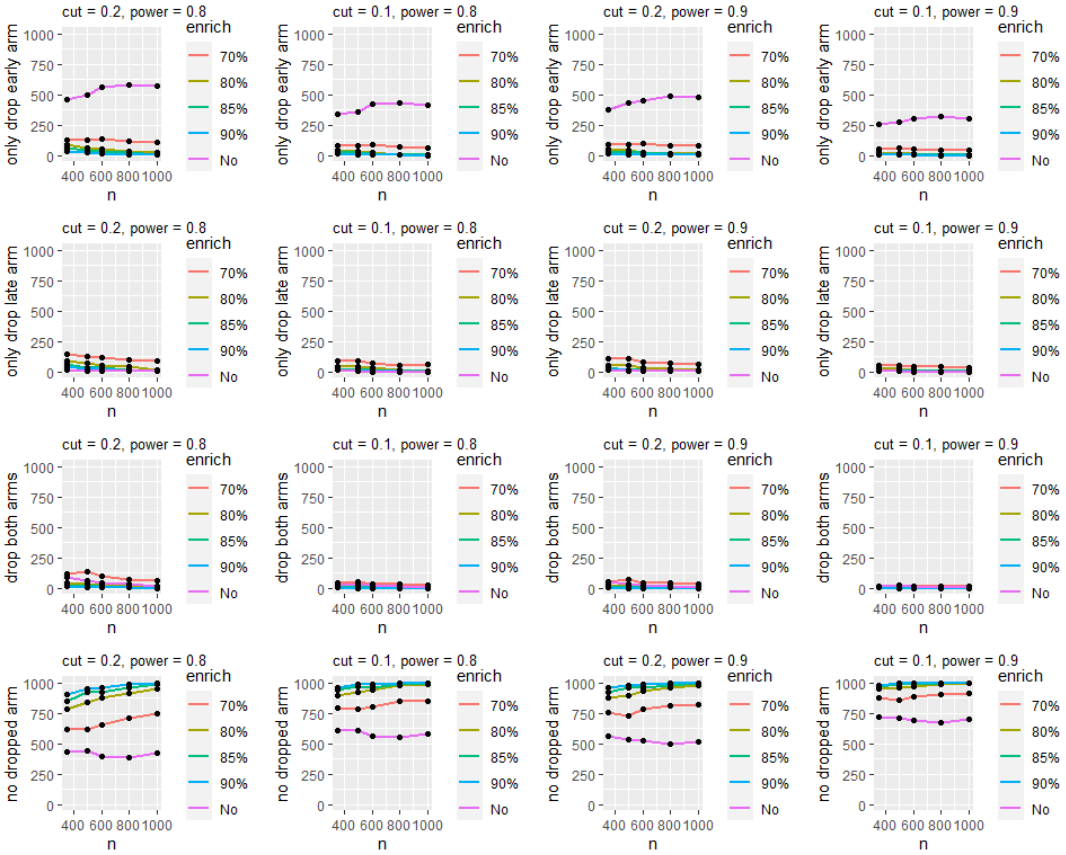


Figure 4.4.3.2 Number of times that early/late surgery group drop at the interim analysis under different settings. The settings include different cutoff point, expected power and the sample size of arm; x-axis denotes sample size of each arm, y-axis denotes the number of dropping times, the title shows the stopping cut point and expected power; "enrich" shows what percentage of patients down to 15cc or less clot is included in the simulated trial; "cut" means arm stopping criteria in the interim analysis

Figure 4.4.3.2 shows the results of 1000 simulated trials under different scenarios and compares the number of dropping times at interim analysis. The first and second rows of figures show the number of times that only dropping early/late surgery group at interim analysis. The third row of the figures shows the number of times that both early and late groups stop at interim analysis. The last row of the figures shows how many times that no surgery arms stop early. Line color shows the number of dropping times when having different enhancements for patients with 15cc or less clot at the end of treatment in the simulated trial, "No" means there is no prespecified enhancement.

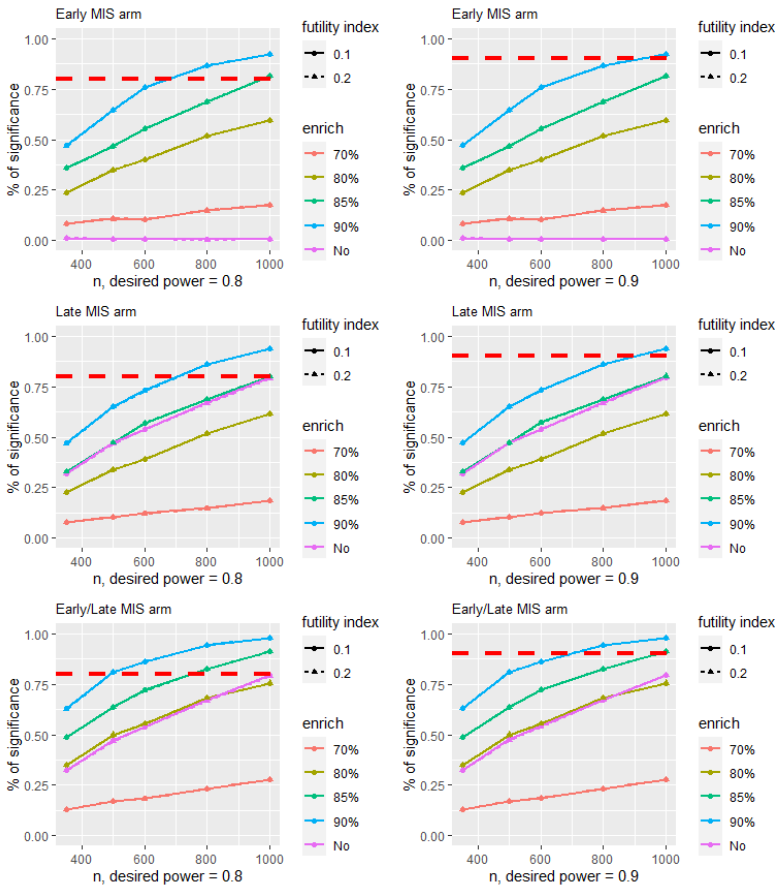


Figure 4.4.3.3 Proportion of simulated trials rejecting corresponding H0; x-axis denotes sample size of each arm; y-axis denotes the proportion of trials rejecting the corresponding H0 of the early MIS arm(first row), the late MIS arm(second row) or at least one MIS arm(third row); futility index 0.1,0.2 denotes the arm stopping criteria at the interim analysis; The red lashed horizontal line denotes desired power; Given the fixed desired power, prespecified enhancement ratio and sample size, powers are pretty closed so that we only can see solid lines in these plots.

Figure 4.4.3.3 compares proportions of simulated trials detecting treatment effect in surgery arms based on different futility index. This figure examines results when we expect to have 80%, 90% power in detecting a clinically important difference between the surgery arm and standard medical arm at the endpoint. Given the fixed desired power and sample size, powers are pretty closed so that we only can see solid lines in these plots. The first and second rows of figures indicate the power of MIS early arm vs. standard medical arm and MIS late arm vs. standard medical arm. The third row of figures shows the power to reject at least one of the surgery arms. In addition, different

colors represent situations with different percentages of patients having 15cc or less clot at the end of treatment in the simulated trials, while "No" means there is no prespecified percentage.

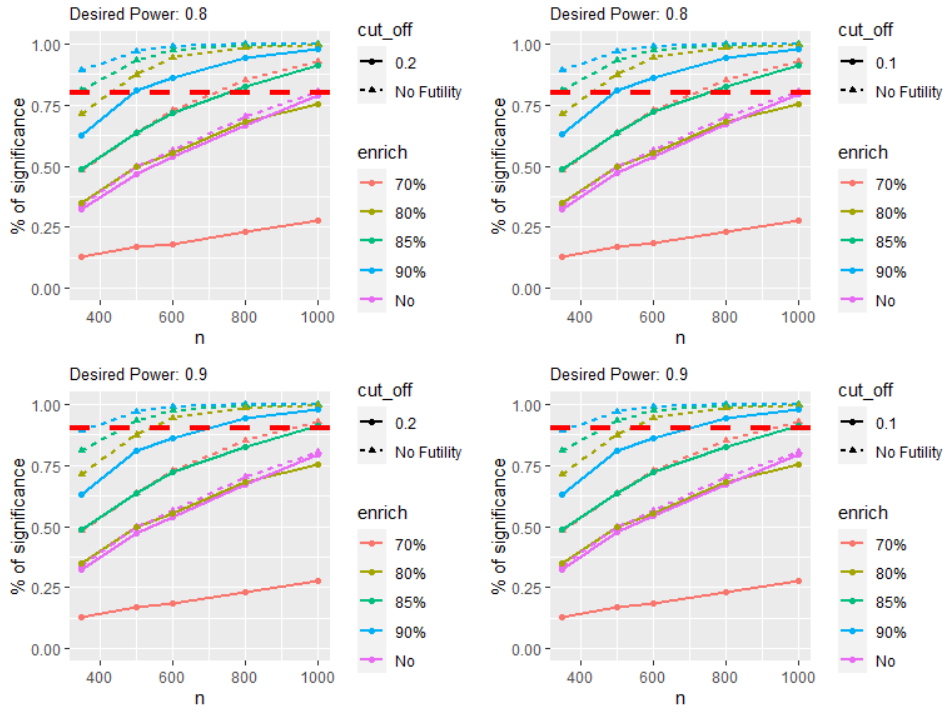


Figure 4.4.3.4 Comparison of power in rejecting at least one MIS arm between cases with and without futility; x-axis denotes sample size of each arm; y-axis denotes the proportion of trials rejecting at least one corresponding H0 of MIS arm; futility index 0.1,0.2 denotes the arm stopping criteria at the interim analysis; "No futility" denotes case without considering futility analysis. The red lashed horizontal line denotes the desired power

Figure 4.4.3.4 compares the proportion of simulated trials detecting at least one MIS surgery arm with a significant result between simulation with and without futility analysis. The calculated power of no futility case can be found in Appendix 3. The first row of the figures shows the comparison when desired power is 0.8 in the setting with futility analysis, while the second row of the figures displays the situation when desired power is 0.9. In addition, different colors represent cases with different percentages of patients having 15cc or less clot at the end of treatment in the simulated trials, while "No" means there is no prespecified percentage.

4.5 Discussion

In Figure 4.4.3.1, we use simulated trials with a sample size of each arm equal to 600 as an example. When having the same expected power and futility stopping criteria, simulation with 90% enhancement of patients achieving 15cc or less clot size has fewer arm drops at the interim analysis. It detects more trials having differences between surgery arm and standard medical arm at the endpoint.

In Figure 4.4.3.2, we should be aware of potential variability in the MIS arms stopping times at interim analysis but still can see a difference between simulated trials with and without enhancement. These figures show that MIS early (evacuation timing < 55hrs) arms usually have fewer stops at the interim analysis when the simulated trials propose enhancement in patients with 15cc or less clot at the end of treatment. If the expected power is 0.8 and interim stopping criteria is 0.2, there are less than approximately 500 MIS early arm stops in simulated 1000 trials with specified enhancement compared to simulated trials with no enhancement. A higher percentage of patients' enhancement in achieving clot size 15cc or less at the end of treatment results in fewer opportunities that surgery arm stops at the interim analysis. We can also see that larger stopping criteria at the interim analysis bring more stops, which is reasonable for more trials to fall below the futility threshold.

Figure 4.4.3.3 compares proportions of simulated trials detecting treatment effects under three settings. (1) MIS early vs. Medical (2) MIS late vs. Medical (3) at least one MIS arm. There is no big difference in results when proposing 0.1/0.2 stopping criteria at the interim analysis. In addition, the larger sample size of the arm brings higher possibilities of detecting treatment effects. In the setting with 90% patients enhancement in achieving clot size less than 15 cc,

simulations with the arm size larger than 600 can reach the desired power in detecting significant outcome for at least one MIS arm when the desired power is 0.8.

Figure 4.4.3.4 compares the proportion of simulated trials detecting treatment effect between simulation with and without futility analysis. Different futility indexes do not have a significant impact in detecting treatment effects. When desired power is 0.9, a few scenarios can reach the desired power. Simulations without futility analysis achieve higher power than simulations with futility analysis, which is reasonable that dropping arms in the simulations with futility analysis can dilute the detectable treatment effects. Also, we should be aware that the impact of the futility stopping rule on power should be small (e.g., at most 1-2% decrease in power). When simulations have arm size = 1000 and prespecify 90% of patients in MIS arm achieving clot size less than 15cc at the end of treatment, the difference of power in-between settings with and without futility analysis is smaller than 2.5%.

As we can see, many factors affect the final outcomes. Thus it is necessary to set up an appropriate expected power and futility in the clinical study plan to balance clinical resources and researchers' expectations.

.

Chapter 5

Discussion and Conclusion

In this thesis work, we first introduce current research in MIS benefits to stroke patients and the necessity in providing power analysis and more rigorous MIS performance for future clinical studies. We compare the models' performance in binary endpoint and ordinal endpoint based on MISTIE II/III. When examining power based on simulation, we find that larger sample size and higher enhancement in patients achieving 15cc or less clot size results in higher desired power. Absolute gains in power vary from 2% to 45%, with sample size ranging from 350 to 1000 participants and enhancement = "No enhancement", 70%, 80%, 85%, and 90%. When the sample size is above 600, desired power can be above 90% no matter the enhancement, which we consider potential resource waste. We then evaluate trial results when taking into account futility and patients' clot evacuation timing. Higher enhancement in MIS patients brings fewer stops at the interim analysis, especially in MIS early surgery arm. Higher enhancement achieves higher power at the endpoints. Also, simulation with enhancement narrows the gap between times of early and late surgery arm dropped.

In summary, to balance desired power and clinical resources, we suggest that in detecting MIS treatment effect for ICH patients compared to standard medical care, future clinical studies should have as many patients as possible achieving 15cc or less clot size at the end of MIS. It is reasonable if the sample size of each arm $\geq$ 800 with power = 80%. If clinicians prefer a smaller

sample size, then analysis of treatment effects with covariate adjustment and ordinal endpoints can be considered.

This study estimates the effect size from the actual MIS data, which could be a potential limitation. Furthermore, the analyses in Chapter 3 and 4 do not take into account covariate adjustment. We recommend considering covariate adjustment in the future analysis. Another limitation is that since clot volume is a post-treatment variable, there could be unmeasured confounding of the effect of achieving a smaller clot volume on the primary outcome (mRS).

Appendix 1

1. Demographic and baseline characteristics for MISTIE II/III clinical trials stratified by patients symptom onset to MIS surgery hours (Late ≥ 55hrs versus Early < 55hrs). "Stability CT ICH" denotes Stability CT Intracerebral hemorrhage(ICH) volume, "Stability CT IVH" denotes Stability CT Intraventricular hemorrhage (IVH) volume, Glasgow Coma Scale (GCS) is a scoring system used to describe patients' degree of consciousness following a brain injury. [11] The clot location is either lobar or deep (putamen or thalamus).

	Late (N=139)	Early (N=135)	Medical (N=280)	Overall (N=554)
age				
Mean (SD)	60.6 (12.3)	61.7 (10.8)	61.4 (12.9)	61.3 (12.3)
Median [Min, Max]	63.0 [29.0, 84.0]	62.0 [38.0, 81.0]	62.0 [28.0, 90.0]	62.0 [28.0, 90.0]
Missing	0 (0%)	7 (5.2%)	0 (0%)	7 (1.3%)
Glasgow Coma Scale				
3-8	40 (28.8%)	37 (27.4%)	75 (26.8%)	152 (27.4%)
9-12	59 (42.4%)	58 (43.0%)	116 (41.4%)	233 (42.1%)
13-15	40 (28.8%)	39 (28.9%)	89 (31.8%)	168 (30.3%)
Missing	0 (0%)	1 (0.7%)	0 (0%)	1 (0.2%)
clot location				
Deep(putamen or thalamus)	87 (62.6%)	94 (69.6%)	166 (59.3%)	347 (62.6%)
Lobar	52 (37.4%)	41 (30.4%)	114 (40.7%)	207 (37.4%)
stability CT ICH				
Mean (SD)	48.1 (18.8)	49.6 (16.8)	48.0 (17.6)	48.4 (17.7)
Median [Min, Max]	43.5 [20.9, 127]	47.7 [20.9, 99.3]	45.2 [16.8, 119]	45.3 [16.8, 127]
stability CT IVH				
Mean (SD)	2.25 (4.87)	3.48 (7.73)	2.68 (4.75)	2.77 (5.66)
Median [Min, Max]	0.200 [0, 42.7]	0.400 [0, 61.8]	0.500 [0, 29.5]	0.400 [0, 61.8]
mRS at 180 days(ordinal)				
0	1 (0.7%)	1 (0.7%)	4 (1.4%)	6 (1.1%)
1	6 (4.3%)	1 (0.7%)	5 (1.8%)	12 (2.2%)
2	16 (11.5%)	16 (11.9%)	25 (8.9%)	57 (10.3%)
3	37 (26.6%)	30 (22.2%)	68 (24.3%)	135 (24.4%)
4	35 (25.2%)	34 (25.2%)	66 (23.6%)	135 (24.4%)
5	25 (18.0%)	27 (20.0%)	43 (15.4%)	95 (17.1%)
6	19 (13.7%)	26 (19.3%)	69 (24.6%)	114 (20.6%)
mRS at 180 days(binary)				
0-3	60 (43.2%)	48 (35.6%)	102 (36.4%)	210 (37.9%)
4-6	79 (56.8%)	87 (64.4%)	178 (63.6%)	344 (62.1%)

"mRS": Patients Modified Rankin Scale(mRS) - ordinal at day 180, range from 0 to 6 where 0 is no symptom, 1 is no significant disability and able to carry out all pre-stroke activities, 2 is slight disability and unable to carry out all pre-stroke activities, 3 is moderate disability with requiring some external help, 4 is moderately severe disability and unable to walk or attend to bodily functions without others' assistance, 5 is severe disability with requiring continuous care, 6 is death; binary at day 180, 1 = good mRS (0-3), 0 = bad mRS (4-6)

Appendix 2

1. Futility results (out of 1000 simulated trials) when expected power is 80% with different arm sizes n = 350, 500, 600, 800, and 1000. Stopping cutoff = 0.2 at interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial.

n	Only drop MIS early arm	Only drop MIS late arm	Drop both MIS arms	Enhancement
350	463	17	92	No
500	496	2	58	No
600	561	2	41	No
800	582	2	31	No
1000	567	1	12	No
350	33	45	18	90%
500	25	24	6	90%
600	17	20	3	90%
800	6	7	1	90%
1000	2	2	0	90%
350	63	61	25	85%
500	38	32	12	85%
600	30	37	10	85%
800	21	17	6	85%
1000	8	5	0	85%
350	93	84	43	80%
500	62	71	29	80%
600	53	50	21	80%
800	37	40	10	80%
1000	23	16	10	80%
350	123	142	118	70%
500	124	125	133	70%
600	138	113	96	70%
800	114	101	74	70%
1000	104	87	62	70%

2. The number of simulated trials that detect a significant difference between surgery arm and standard medical arm given the different stopping scenarios. The expected power is 80%, arm sizes n = 350, 500, 600, 800, and 1000, stopping cutoff = 0.2 at the interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial

n	Only drop MIS early arm	Only drop MIS late arm	Both MIS arms remain				Enhancement
			Early vs. Medical	Late vs. Medical	Both	Total	
350	118/463	0/17	2	195	8	205/428	No
500	196/496	0/2	1	268	4	273/444	No
600	278/561	0/2	1	259	3	263/396	No
800	356/582	0/2	0	309	3	312/385	No
1000	421/567	0/1	0	363	7	370/420	No
350	7/33	11/45	147	151	311	609/904	90%
500	13/25	7/24	151	150	487	788/945	90%
600	7/17	7/20	123	95	628	846/960	90%
800	3/6	4/7	76	74	785	935/986	90%
1000	1/2	2/2	39	53	882	974/996	90%
350	9/63	13/61	145	119	199	463/851	85%
500	9/38	8/32	154	160	303	617/918	85%
600	7/30	12/37	141	157	403	701/923	85%
800	9/21	7/17	130	125	552	807/956	85%
1000	4/8	0/5	110	93	704	907/987	85%
350	5/93	12/84	112	109	112	333/780	80%
500	11/62	16/71	146	140	187	473/838	80%
600	8/53	8/50	159	147	233	539/876	80%
800	10/37	10/40	154	153	353	660/913	80%
1000	9/23	6/16	137	152	452	741/951	80%
350	6/123	5/142	48	42	28	118/617	70%
500	6/124	8/125	61	58	37	156/618	70%
600	4/138	8/113	53	74	43	170/653	70%
800	5/114	8/101	76	78	63	217/711	70%
1000	7/104	5/87	90	98	78	266/747	70%

The table's heading denotes different situations at the interim analysis, "NA" denotes that the surgery arm was not dropped in any of the 1000 simulations for this scenario. For example, 118/463 means 188 trials detect a significant difference in the endpoint given 463 trials having MIS early arm dropped at the interim analysis

3. Futility results (out of 1000 simulated trials) when expected power is 80% with different arm sizes n = 350, 500, 600, 800, and 1000. Stopping cutoff = 0.1 at interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial.

n	Only Drop MIS early arm	Only Drop MIS late arm	Drop both MIS arms	Enhancement
350	343	16	34	No
500	361	2	26	No
600	421	0	21	No
800	435	0	10	No
1000	417	0	2	No
350	14	18	7	90%
500	3	9	0	90%
600	6	6	0	90%
800	2	1	1	90%
1000	0	0	0	90%
350	28	24	8	85%
500	20	11	1	85%
600	18	15	5	85%
800	8	6	0	85%
1000	2	1	0	85%
350	43	43	18	80%
500	29	42	4	80%
600	21	29	11	80%
800	8	15	4	80%
1000	8	7	1	80%
350	79	91	41	70%
500	80	84	51	70%
600	86	74	37	70%
800	70	54	29	70%
1000	61	65	26	70%

4. The number of simulated trials that detect a significant difference between surgery arm and standard medical arm given the different stopping scenarios. The expected power is 80%, arm sizes n = 350, 500, 600, 800, and 1000, stopping cutoff = 0.1 at the interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial

n	Only drop MIS early arm	Only drop MIS late arm	Both MIS arms remain				Enhancement
			Early vs. Medical	Late vs. Medical	Both	Total	
350	70/343	0/16	3	243	8	254/607	No
500	123/361	0/2	1	344	4	349/611	No
600	189/421	NA	1	348	3	352/558	No
800	240/435	NA	1	426	4	431/555	No
1000	290/417	NA	0	496	7	503/581	No
350	4/14	5/18	154	155	311	620/961	90%
500	1/3	2/9	156	160	490	806/988	90%
600	0/6	1/6	128	101	630	859/988	90%
800	0/2	0/1	80	77	785	942/996	90%
1000	NA	NA	41	54	882	977/1000	90%
350	2/28	3/24	154	127	200	481/940	85%
500	4/20	3/11	161	166	303	630/968	85%
600	5/18	3/15	149	162	404	715/962	85%
800	0/8	1/6	135	134	553	822/986	85%
1000	0/2	0/1	110	97	705	912/997	85%
350	2/43	1/43	122	111	114	347/896	80%
500	4/29	7/42	155	147	187	489/925	80%
600	1/21	6/29	162	154	233	549/939	80%
800	1/8	3/15	162	163	353	678/973	80%
1000	4/8	1/7	142	156	453	751/984	80%
350	1/79	1/91	52	47	28	127/789	70%
500	2/80	4/84	64	62	38	164/785	70%
600	1/86	3/74	57	78	44	179/803	70%
800	3/70	3/54	81	81	63	225/847	70%
1000	2/61	1/65	94	103	78	275/848	70%

The table's heading denotes different situations at the interim analysis, "NA" denotes that the surgery arm was not dropped in any of the 1000 simulations for this scenario. For example, 70/343 means 70 trials detect a significant difference in the endpoint given 343 trials having MIS early arm dropped at the interim analysis

5. Futility results (out of 1000 simulated trials) when expected power is 90% with different arm sizes n = 350, 500, 600, 800, and 1000. Stopping cutoff = 0.2 at interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial.

n	Only Drop MIS early arm	Only Drop MIS late arm	Drop both MIS early arm	Enhancement
350	373	16	49	No
500	432	2	31	No
600	454	1	24	No
800	486	1	14	No
1000	476	1	5	No
350	16	21	8	90%
500	6	13	0	90%
600	7	9	0	90%
800	2	3	1	90%
1000	2	1	0	90%
350	35	29	12	85%
500	22	17	5	85%
600	22	18	5	85%
800	12	9	1	85%
1000	3	2	0	85%
350	51	49	25	80%
500	40	53	14	80%
600	23	34	15	80%
800	13	23	5	80%
1000	12	11	3	80%
350	88	104	56	70%
500	91	110	70	70%
600	98	78	45	70%
800	80	71	39	70%
1000	75	65	37	70%

6. The number of simulated trials that detect a significant difference between surgery arm and standard medical arm given the different stopping scenarios. The expected power is 90%, arm sizes n = 350, 500, 600, 800, and 1000, stopping cutoff = 0.2 at the interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial

n	Only drop MIS early arm	Only drop MIS late arm	Both MIS arms remain				Enhancement
			Early vs. Medical	Late vs. Medical	Both	Total	
350	85/373	0/16	3	228	8	239/562	No
500	158/432	0/2	1	309	4	314/535	No
600	210/454	0/1	1	327	3	331/521	No
800	272/486	0/1	1	393	4	398/499	No
1000	340/476	0/1	0	446	7	453/518	No
350	4/16	5/21	154	155	311	620/955	90%
500	2/6	5/13	154	159	489	802/981	90%
600	0/7	3/9	126	101	630	857/984	90%
800	0/2	2/3	78	77	785	940/994	90%
1000	1/2	1/1	40	53	882	975/997	90%
350	3/35	3/29	154	125	200	479/924	85%
500	3/22	4/17	158	166	303	627/956	85%
600	6/22	3/18	149	160	404	713/955	85%
800	3/12	2/9	134	131	553	818/978	85%
1000	1/3	0/2	110	97	704	911/995	85%
350	3/51	3/49	120	110	114	344/875	80%
500	6/40	10/53	152	145	187	484/893	80%
600	1/23	6/34	161	154	233	548/928	80%
800	3/13	4/23	161	161	353	675/959	80%
1000	4/12	4/11	139	156	453	748/974	80%
350	2/88	2/104	51	46	28	125/752	70%
500	5/91	6/110	63	59	37	159/729	70%
600	3/98	4/78	56	76	44	176/779	70%
800	3/80	3/71	81	81	63	225/810	70%
1000	4/75	1/65	94	101	78	273/823	70%

The table's heading denotes different situations at the interim analysis, "NA" denotes that the surgery arm was not dropped in any of the 1000 simulations for this scenario. For example, 85/373 means 85 trials detect a significant difference in the endpoint given 373 trials having MIS early arm dropped at the interim analysis

7. Futility results (out of 1000 simulated trials) when expected power is 90% with different arm sizes n = 350, 500, 600, 800, and 1000. Stopping cutoff = 0.1 at interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial.

n	Only Drop MIS early arm	Only Drop MIS late arm	Drop both MIS early arm	Enhancement
350	255	10	14	No
500	273	2	16	No
600	301	0	12	No
800	321	1	4	No
1000	300	0	2	No
350	9	8	2	90%
500	1	7	0	90%
600	5	2	0	90%
800	0	0	0	90%
1000	0	0	0	90%
350	16	15	1	85%
500	9	4	0	85%
600	6	9	3	85%
800	5	2	0	85%
1000	2	1	0	85%
350	24	23	5	80%
500	14	25	1	80%
600	12	17	3	80%
800	5	8	0	80%
1000	3	4	0	80%
350	53	55	13	70%
500	65	53	25	70%
600	56	42	16	70%
800	41	45	12	70%
1000	41	36	11	70%

8. The number of simulated trials that detect a significant difference between surgery arm and standard medical arm given the different stopping scenarios. The expected power is 90%, arm sizes n = 350, 500, 600, 800, and 1000, stopping cutoff = 0.1 at the interim analysis, with no enhancement or having enhancement = 90%, 85%, 80%, 70% of patients down to 15cc or less clot at the end of surgery in each simulated trial

n	Only drop MIS early arm	Only drop MIS late arm	Both MIS arms remain				Enhancement
			Early vs. Medical	Late vs. Medical	Both	Total	
350	46/255	0/10	4	266	9	279/721	No
500	85/273	0/2	4	382	4	390/709	No
600	122/301	NA	3	414	4	421/687	No
800	174/321	0/1	1	492	4	497/674	No
1000	196/300	NA	0	590	7	597/698	No
350	1/9	2/8	157	158	311	626/981	90%
500	0/1	2/7	156	161	490	807/992	90%
600	0/5	0/2	129	101	630	860/993	90%
800	NA	NA	81	77	785	943/1000	90%
1000	NA	NA	41	54	882	977/1000	90%
350	1/16	2/15	155	128	200	483/968	85%
500	3/9	0/4	164	167	303	634/987	85%
600	1/6	2/9	150	167	404	721/982	85%
800	0/5	0/2	136	134	553	823/993	85%
1000	0/2	0/1	110	97	705	912/997	85%
350	1/24	0/23	123	112	114	349/948	80%
500	1/14	4/25	158	150	187	495/960	80%
600	0/12	3/17	165	155	233	553/968	80%
800	0/5	2/8	163	164	353	680/987	80%
1000	1/3	0/4	143	159	453	755/993	80%
350	1/53	0/55	53	47	28	128/879	70%
500	1/65	1/53	67	63	38	168/857	70%
600	1/56	1/42	59	78	44	181/886	70%
800	1/41	2/45	82	83	63	228/902	70%
1000	0/41	0/36	95	106	78	279/912	70%

The table's heading denotes different situations at the interim analysis, "NA" denotes that the surgery arm was not dropped in any of the 1000 simulations for this scenario. For example, 46/255 means 46 trials detect a significant difference in the endpoint given 255 trials having MIS early arm dropped at the interim analysis

Appendix 3

1. The proportion of simulated trials detecting at least one MIS arm is significant under no futility settings and different enhancement ratio

N	No Enhancement	70% Enhancement	80% Enhancement	85% Enhancement	90% Enhancement
350	0.341	0.483	0.712	0.81	0.893
500	0.497	0.638	0.874	0.935	0.969
600	0.566	0.728	0.942	0.973	0.991
800	0.702	0.85	0.982	0.996	0.999
1000	0.804	0.927	0.992	1	1

N denotes simulated arm size for the MIS and the standard medical groups; “Enhancement” denotes the proportion of patients having 15cc or less clot at the end of treatment in the simulated trials

Bibliography

1. Qureshi AI, Tuhrim S, Broderick JP, Batjer HH, Hondo H, Hanley DF. Spontaneous intracerebral hemorrhage. *N Engl J Med.* 2001;344:1450–60.
2. Sudlow CL, Warlow CP. Comparable studies of the incidence of stroke and its pathological types: results from an international collaboration. International Stroke Incidence Collaboration. *Stroke*. 1997;28(3):491-499. doi:10.1161/01.str.28.3.491
3. Rymer MM. Hemorrhagic stroke: intracerebral hemorrhage. *Mo Med*. 2011;108(1):50-54.
4. Hachinski V, Donnan GA, Gorelick PB, et al. Stroke: working toward a prioritized world agenda. Int J Stroke 2010; 5: 238–56.
5. Mayer SA, Brun NC, Begtrup K, et al. Efficacy and safety of recombinant activated factor VII for acute intracerebral hemorrhage.N Engl J Med 2008; 358: 2127–37.
6. RANKIN J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scott Med J. 1957 May;2(5):200-15. doi: 10.1177/003693305700200504. PMID: 13432835.
7. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. Int J Stroke. 2009;4(3):200-205. doi:10.1111/j.1747-4949.2009.00271.
8. Lees KR, Bath PM, Schellinger PD, et al. Contemporary outcome measures in acute stroke research: choice of primary outcome measure. Stroke. 2012;43(4):1163-1170. doi:10.1161/STROKEAHA.111.641423
9. Safety and efficacy of minimally invasive surgery plus recombinant tissue plasminogen activator in intracerebral hemorrhage evacuation (MISTIE): a randomised, controlled, open-label, phase 2 trial. *Lancet Neurol*. 2016;15(12):1228-1237. doi:10.1016/S1474-4422(16)30234-4

10. Hanley DF, Thompson RE, Rosenblum M, et al. Efficacy and safety of minimally invasive surgery with thrombolysis in intracerebral haemorrhage evacuation (MISTIE III): a randomised, controlled, open-label, blinded endpoint phase 3 trial [published correction appears in Lancet. 2019 Apr 20;393(10181):1596]. Lancet. 2019;393(10175):1021-1032. doi:10.1016/S0140-6736(19)30195-3
11. Awad IA, Polster SP, Carrión-Penagos J, et al. Surgical Performance Determines Functional Outcome Benefit in the Minimally Invasive Surgery Plus Recombinant Tissue Plasminogen Activator for Intracerebral Hemorrhage Evacuation (MISTIE) Procedure. Neurosurgery. 2019;84(6):1157-1168. doi:10.1093/neuros/nyz077
12. Scaggiante J, Zhang X, Mocco J, Kellner CP. Minimally Invasive Surgery for Intracerebral Hemorrhage. Stroke. 2018;49(11):2612-2620. doi:10.1161/STROKEAHA.118.020688
13. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;2(7872):81-84. doi:10.1016/s0140-6736(74)91639-0
14. Jitlal, M., Khan, I., Lee, S. et al. Stopping clinical trials early for futility: retrospective analysis of several randomised clinical studies. Br J Cancer 107, 910–917 (2012). https://doi.org/10.1038/bjc.2012.344
15. Jennison, Christopher, and Bruce W. Turnbull. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman & Hall/CRC, 2000.
16. Chang, Mark. Classical and Adaptive Clinical Trial Designs. John Wiley & Sons. Hoboken, New Jersey, 2008
17. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946-1978. doi:10.1002/sim.6082

CURRICULUM VITAE

Yuying Yan

615 N. Wolfe Street
Baltimore, MD 21205
+1(608)-320-0020; yyan42@jhu.edu

EDUCATION BACKGROUND

The Johns Hopkins University - Bloomberg School of Public Health September 2019 - May 2021
M.S. Biostatistics GPA:3.94/4.0

- Selected Courses: Methods in Biostatistics, Probability Theory, Advanced Data Science, Survival Analysis, Risk Prediction and Precision Medicine, Analysis of Longitudinal Data, Design of Clinical Experiments, Introduction to Data Management

University of Wisconsin-Madison | Graduate with Distinction, *Phi Beta Kappa* September 2015 - May 2019
B.S. in Mathematics, Statistics GPA: 3.87/4.0

- Selected Courses: Computational Statistics, Statistical Experimental Design, Introduction to Bioinformatics, Stochastic Process, Introduction to Data Structures, Statistical Analysis in Categorical Variables

SKILLS

Proficient in SAS, R, SQL, Python
Intermediate in Java, MATLAB, STATA, REDCap, Github
Certificate in Base Programming Using SAS 9.4

RESEARCH EXPERIENCE

Biostatistics Research Assistant |September 2020 - current
Advisor: Richard E Thompson Baltimore, MD

- Project I: Compute **Cumulative Incidence function** for Intracerebral Hemorrhage patients' length of hospitalization with taking account death as competing risk by R and Implement **Cox regression model** to inspect covariates that have significant impact on length of hospitalization.
- Project II: **Simulate** power and inspect multiple ways to adjust alpha such as **Hochberg approach** to avoid overpowered for the proposed MISTIE (minimally invasive surgery with thrombolysis in intracerebral Hemorrhage evacuation) IV Trial in R language

Undergraduate Researcher | February 2018 - August 2018
Advisor: Qiongshi Lu Madison, WI

- Adjusted **Congenital Heart Disease empirical Bayesian framework** for gene variant enrichment analysis to increase statistically power based on some gene-related scores
- Used empirical Bayesian framework calculate and adjusted theoretical gene frequency by **gene relevant scores (pLI score)**, compared theoretical frequency with observed gene frequency in Congenital Heart Disease patients
- Brought provision to biologists in identifying probable Congenital Heart Disease associated gene

WORK EXPERIENCE

Internship – Management Consulting | January 2021 - current
IQVIA (Shanghai) Remote

- Help clients analyze important factors in enrolling National Reimbursement Drug List and interview with Chinese e34 experts
- Conduct literature review and organize drugs relevant clinical research files

Volunteer | January 2021 - current
Johns Hopkins Biostatistics center Baltimore, MD

- Shadow ICTR clinics with biostatistics faculty in assisting researcher with study design, statistical analysis, data interpretation, etc.

TEACHING EXPERIENCE

Graduate Teaching Assistant | June 2020 - current
Johns Hopkins University Department of Biostatistics Baltimore, MD

- Develop clear content and respond to student questions for the course: Statistical Methods in Public Health
- Collaborate with team of teaching assistants and faculties to actively design new ideas on courses and contribute to the academic success of students

Course Grader - Introduction to Probability Theory| September 2018 – May 2019
University of Wisconsin-Madison, Department of Mathematics Madison, WI

- Graded students' homework in a weekly basis and helped Lecturer organize students grade

Course Assistant - Combinatorics | September 2018 – December 2018
University of Wisconsin-Madison, Department of Mathematics Madison, WI

- Developed clear content in helping students understand course content and solve homework questions in a weekly basis.

Tutor - Calculus/Statistics | June 2017 – December 2017
University of Wisconsin-Madison, DDEEA Madison, WI

- Helped minority group students strengthen comprehension in their mathematics/statistics courses
- Supported the mission to create a diverse, inclusive and excellent learning environment for all students at the university

Honors/Awards

Dean's List
University of Wisconsin - Madison Madison, WI

- Received Dean's List for 5 semesters

EXTRACURRICULUM

COVID-19 Data Collection Volunteer | March 2020 - May 2020
Johns Hopkins University Baltimore, MD

- Collected Laos government's response to COVID-19 by finding local news and government documents
- Formalized timeline of government's response to accelerate policy research progress

Donation
Dane County Humane Society Madison, WI

- Donate cat food and necessities every six months since 2018